Active Learning of Spin Network Models

Jialong Jiang[a,1], **David A. Sivak**[b], **and Matt Thomson**[a,1]

[a]Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA 91125; [b]Department of Physics, Simon Fraser University, Burnaby, BC V5A 1S6, Canada

This manuscript was compiled on May 15, 2019

The inverse statistical problem of finding direct interactions in complex networks is difficult. In the context of the experimental sciences, well-controlled perturbations can be applied to a system, probing the internal structure of the network. Therefore, we propose a general mathematical framework to study inference with iteratively applied perturbations to a network. Formulating active learning in the language of information geometry, our framework quantifies the difficulty of inference as well as the information gain due to perturbations through the curvature of the underlying parameter manifold as measured though the empirical Fisher information. Perturbations are then chosen that reduce most the variance of the Bayesian posterior. We apply this framework to a specific probabilistic graphical model where the nodes in the network are modeled as binary variables, "spins" with Ising-form pairwise interactions. Based on this strategy, we significantly improve the accuracy and efficiency of inference from a reasonable number of experimental queries for medium sized networks. Our active learning framework could be powerful in the analysis of complex networks as well as in the rational design of experiments.

Network | Inference | Active Learning | Information Geometry

A significant property of complex systems is the convoluted interaction between different parts. Describing the structure of interactions in the network is critical to understanding and predicting its behavior. Numerous models have been developed to characterize complex networks, and many different methods are used to infer network interactions from the data generated by a network, for example, methods based on variable correlation, mutual information between variables, likelihood, and temporal dynamic relationships (1).

However, difficulties are always confronted while solving the problem of deducing direct interaction from correlation, as many alternative causal relations all can explain the same observed correlations. Many disciplines in scientific research, especially biology, rely on perturbations to tackle the inference problem. For example, gene functions are studied by their mutants, and signaling pathways are decoded from carefully designed knock-in/out experiments. Recent developments in molecular biology provide high-throughput technology to perform perturbation experiments, such as CRISPR/Cas9 in gene editing (2, 3) and optogenetics in neuron activity control (4). With these methods, it is natural to ask how to design perturbation experiments and extract information from the data to make the best possible inference.

There have been studies of optimal design (5, 6) and analysis (7) of perturbation experiments, and efforts to connect the two in an iterative process. Boolean networks were studied, for example, when there were only population average data available (8). Moreover, active learning of Bayesian networks on directed acyclic graphs has been studied from many facets in causal inference. Interventions are modeled as pinning down node values to distinguish between Markov equivalent models (9–12). Nonetheless, previous works address specific classes of models, and a general mathematical framework to understand how perturbations facilitate inference is still absent.

To formulate the problem and demonstrate the framework, we constrain ourselves to a specific probabilistic graphical model of complex networks: spin networks. Nonetheless, the framework developed in this paper tries to capture a ubiquitous structure in physical inference problems and can be generalized to other probabilistic models without difficulty. Among modeling approaches, spin network models are one of the simplest and most natural in the sense that such models provide the maximum entropy inference of a network given the means and correlations of nodes in the network (13). The inference problem involved in parametrizing a spin network model from data is known in physics as the inverse Ising problem (or spin glass inverse problem), and has been widely applied to many fields such as computational biology (14, 15), neuroscience (16), data science (17), and so on (13).

A spin network is a probabilistic graphical model with each node taking value in $\{1, -1\}$. For a p-node network, the probability distribution over 2^p configurations is the induced Boltzmann distribution from an Ising-type interaction energy. Then the probability of a configuration $\boldsymbol{s}$ given an interaction matrix $\boldsymbol{J}$ and local field $\boldsymbol{h}$ is

$$P(\boldsymbol{s}|\boldsymbol{J},\boldsymbol{h}) = \frac{\exp[-E(\boldsymbol{s})]}{\mathcal{Z}}\,,$$
$$E(\boldsymbol{s}) = -\sum_{i<j} J_{ij}s_is_j - \sum_i h_is_i\,, \qquad [1]$$

Significance Statement

The inverse problem of learning direct interaction from correlation is always difficult. In the natural sciences, applying interventions is crucial to uncover such interactions. The development of new technologies now enables the performance of high-throughput perturbation experiments to probe network structure. However, the analysis and design of such experiments still remains mostly empirical. We develop an inference framework that allows automated and active learning of statistical models through iterative rounds of observation and perturbation. Fisher information is used to quantify the uncertainty of inference, and to design new perturbations. The learning process actively explores the behavior of a given system and combines all information in a Bayesian framework. Our problem formulation also raises new statistical problems for further investigation.

M.T. and D.S. conceived of the idea. J.J. and M.T. developed the theory and performed the computations. J.J. developed the Bayesian inference framework. J.J. and M.T. wrote the manuscript.

The authors declare no conflict of interest.

[1]To whom correspondence should be addressed. E-mail: mthomson@caltech.edu or jiangjl@caltech.edu

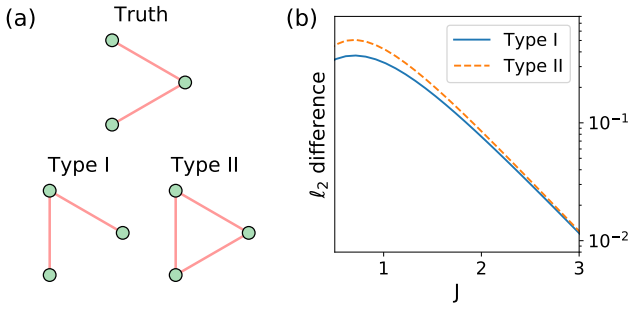

Fig. 1. The difficulty of inferring a three-node network. (a) During network inference, two common types of error could occur. Type I error occurs when some correct edges are missing and the inference may contain incorrect edges in compensation. This is because different interaction topologies may result in a similar correlation pattern in data. Type II error occurs when the inferred network includes extra incorrect interaction edges, which is related to correlations between nodes caused by indirect interactions. (b) The spin-spin correlations in data generated by the three different networks from (a) with high interaction strengths are nearly identical. The ℓ_2 difference of spin-spin correlation vectors caused by the two types of incorrect structures is plotted as a function of J, where all interaction strengths are set to J. The ℓ_2 difference decreases exponentially with J, so distinguishing different structures is extremely difficult with finite sampling.

where $\mathcal{Z} = \sum_{\{\boldsymbol{s}\}} \exp[-E(\boldsymbol{s})]$ is the partition function. For simplicity, the inverse temperature factor β is absorbed into the parameters for interactions and field strengths. The learning or inference of the model consists of finding the best $\boldsymbol{J}$ and $\boldsymbol{h}$ to describe the observed data, which can be solved by maximum-likelihood estimation (MLE), pseudo-likelihood (18) or other approximate optimization methods (19).

To solve the inverse Ising problem is hard both in sampling complexity and computational cost (20, 21). Even though solving for MLE is a convex optimization problem, the Hessian matrix can be close to singular (22), making it difficult to distinguish alternatives of parameter values. A canonical demonstration is a three-node network whose three nodes are all strongly correlated, as shown in Fig. 1 (a). All interactions have the same strength $J > 0$, so the correlation between any pair of nodes is close to 1 for large J. The difference in correlation caused by different network structures decreases exponentially with the coupling strength, as shown in Fig. 1 (b). Considering that the correlations and means are sufficient statistics of the problem, it is almost impossible to find the correct structure without extra information. Formally, detailed analysis of sampling complexity shows that the number of samples needed to distinguish different structures grows polynomially with the number of edges, but exponentially with the ℓ_∞ norm of $\boldsymbol{J}$, which represents the coupling strength between nodes (20). Further, specific examples can be constructed to show that any algorithm acting on observations of correlations alone has a high probability to fail for some networks (21).

In this paper, we propose a framework to perform parametric estimation with the ability to perturb the system so that we can iteratively update our knowledge through different perturbation experiments. In the context of the inverse Ising problem, we learn the coupling matrix $\boldsymbol{J}$ while controlling the field term $\boldsymbol{h}$. We demonstrate procedures for designing experiments and a learning process to achieve significant improvement in inference accuracy on medium-sized networks with strong couplings. This method provides new insights into the spin network model, and can be applied to complex networks in real systems.

Formulation of Inference with Perturbations

The most common perturbations applicable in practice are individually activating/deactivating different nodes, such as knockdown of genes, induced activation of neurons, etc. In a spin network, the local field $\boldsymbol{h}$ describes a tendency of activation for every node. Specifically,

$$\left.\frac{P(s_i=1)}{P(s_i=-1)}\right|_{h_i=h} = \left.\frac{P(s_i=1)}{P(s_i=-1)}\right|_{h_i=0} \exp(2h)\,. \qquad [2]$$

Therefore, it is natural to consider a scenario where we are able to control $\boldsymbol{h}$ to facilitate the inference of $\boldsymbol{J}$. For simplicity, we assume that we have full control of the field, namely the system does not have an unknown intrinsic field. The case with an intrinsic field can be dealt with similarly using our framework.

Quantification of the difficulty of inference is necessary to design a field that alleviates it, and information geometry provides such a measure. Information geometry defines a geometric structure to characterize the change in a probability distribution with changes in underlying parameters. For a parametric family of distributions $P(\boldsymbol{x}|\boldsymbol{\theta})$, the difference between any two distributions measured by Kullback-Leibler divergence can be expanded as a series of the differential parameters change $\delta\boldsymbol{\theta}$

$$\begin{aligned}\mathrm{KL}(P(\boldsymbol{x}|\boldsymbol{\theta}), P(\boldsymbol{x}|\boldsymbol{\theta}+\delta\boldsymbol{\theta})) &= \frac{1}{2}\delta\boldsymbol{\theta}^T\mathcal{I}\delta\boldsymbol{\theta} + \mathcal{O}(\delta\boldsymbol{\theta}^3)\\ \mathcal{I} = -\left\langle \frac{\partial^2 \log P(\boldsymbol{x}|\boldsymbol{\theta})}{\partial\theta_i\partial\theta_j}\right\rangle &= \left\langle \frac{\partial \log P}{\partial\theta_i}\frac{\partial \log P}{\partial\theta_j}\right\rangle\,. \end{aligned} \qquad [3]$$

The Fisher information matrix (FI) $\mathcal{I}$ describes how the parametric density manifold curves. For independent samples, FI is additive. Small FI corresponds to a small change in the probability distribution given a change in parameter values, making the inference difficult. This phenomenon is characterized by the Cramér–Rao bound which states that the covariance $C \succeq \mathcal{I}^{-1}$ for any unbiased estimator, in the sense of $C - \mathcal{I}^{-1}$ being positive semidefinite. One corollary is that $\Omega(\epsilon^{-1}\lambda^{-1})$ samples are needed to achieve error ϵ in expectation on the projection of the parameters onto the eigenvector of FI with eigenvalue λ. On the other hand, FI is also the expectation of the Hessian matrix of the log-likelihood function, representing the difficulty of numerical optimization of the likelihood function.

As the covariance of an estimator is related to the ℓ_2 error of estimation, the following constrained optimization problem

$$\begin{aligned}\min_{\boldsymbol{h}} \quad & \mathbb{E}(\|\boldsymbol{J}-\tilde{\boldsymbol{J}}\|_2) \\ \text{s.t.} \quad & \boldsymbol{s}\sim P(\boldsymbol{s}|\boldsymbol{J},\boldsymbol{h}) \\ & \tilde{\boldsymbol{J}} = \operatorname*{argmax}_{\boldsymbol{J}'} \log P(\{\boldsymbol{s}\}|\boldsymbol{J}',\boldsymbol{h})\end{aligned} \qquad [4]$$

can be tackled by minimizing an asymptotic lower bound, the trace of the inverse of the FI using the applied field $\boldsymbol{h}$

$$\min_{\boldsymbol{h}} \quad \mathrm{Tr}\ \mathcal{I}(P(\boldsymbol{s}|\boldsymbol{J},\boldsymbol{h}))^{-1}\,. \qquad [5]$$

For inverse Ising inference, the FI can be derived from properties of the exponential family of distributions

$$I_{\{ij\},\{kl\}} = \langle s_i s_j s_k s_l \rangle - \langle s_i s_j \rangle \langle s_k s_l \rangle \ , \qquad [6]$$

where $\{ij\}$ corresponds to interaction term J_{ij}. For the diagonal terms, $I_{\{ij\},\{ij\}} = 1 - \langle s_i s_j \rangle^2$, the maximum is achieved when the correlation between two spins s_i, s_j is close to 0. There is a lower bound of $\mathrm{Tr}\, \mathcal{I}^{-1}$ given by

$$\mathrm{Tr}\ \mathcal{I}^{-1} \geq \frac{p^2}{\mathrm{Tr}\ \mathcal{I}} \geq p \,, \qquad [7]$$

which would be achieved when all configurations have the same probability in the p-node network. The optimal value of $\mathrm{Tr}\, \mathcal{I}^{-1}$ can also be achieved by other distributions, and the existence of an $\boldsymbol{h}$ producing such a distribution depends on the structure of the network. In some cases, suitable external fields can be analytically or numerically solved.

For the simplest case, FI of two-spin inference is a scalar, so the optimum is achieved at the maximum of $\mathcal{I}$, where $\langle s_1 s_2 \rangle = 0$. Without the field ($\boldsymbol{h} = 0$) , $\mathcal{I} = 1 - \tanh^2 J = 1/\cosh^2 J$, which decays exponentially with J. The solution of $\langle s_i s_j \rangle = 0$ is (*SI Appendix*, Supplementary text)

$$h_2 = \frac{1}{2} \log \frac{1 - \exp(2J + 2h_1)}{\exp 2J - \exp 2h_1} \,. \qquad [8]$$

The optimal $\mathcal{I} = 1$ can be achieved by an infinite number of (h_1, h_2), and one special approximate solution is $h_1 = -h_2 = J + \log\sqrt{2}$ for large positive J. By introducing the field, FI is increased by a factor exponential in J, which means the sampling complexity is reduced exponentially.

Fig. 2 (a) shows the landscape of $\mathcal{I}$ as a function of h_1, h_2 for $J = 1$. Note that this landscape is nonconvex and the maximum is not unique. The point without field ($h_1 = 0, h_2 = 0$) is a saddle point, with two principal-axis directions $(1, 1)$ and $(1, -1)$, and $\mathcal{I}$ along these two directions are shown in Fig. 2 (b). Intuitively, the difficulty of inference is caused by the high probability of ground-state configurations and the corresponding diminishing probability of higher energy excited states. Fields in the direction of $(1, -1)$ make one of the previous high energy states more accessible, thus increasing FI. On the other hand, the direction of $(1, 1)$ makes the distribution more concentrated on one state, and the FI decreases, thus $(h_1 = 0, h_2 = 0)$ forms a saddle point.

Another canonical model is the finite ferromagnetic Ising chain with periodic boundary conditions, namely an Ising ring. For an analytical solution, we restrict ourselves to the case of knowing the chain structure and inferring the magnitude of individual interaction strengths. The energy function is

$$E = -\sum_{i=1}^{p} J_i s_i s_{i+1} - \sum_{i=1}^{p} h_i s_i \,, \qquad [9]$$

where the convention of $s_{p+1} \equiv s_1$ is used. The FI can be solved approximately when $J_i > 0$ are all equal to J, and $h_i = 0$ (*SI Appendix*, Supplementary text)

$$\mathcal{I}_{\{i,i+1\}\{j,j+1\}} = \begin{cases} 4(p-1)\exp(-4J) & i = j \\ 4\exp(-4J) & i \neq j \end{cases} \qquad [10]$$

The FI is a circular matrix so its eigenvectors have the form $(1, \omega_j, \ldots, \omega_j^{p-1})$, where $\omega_j = \exp(j\ 2\pi i/p)$. There is one larger eigenvalue $\lambda_1 = 8(p-1)\exp(-4J)$ and $(p-1)$ degenerate small eigenvalues $\lambda_2 = 4(p-2)\exp(-4J)$, as shown in Fig. 2 (c). By the symmetry of the system and motivation from the two-node case, a possible good perturbation $h^{(1)}$ can be chosen as $h_j^{(1)} = h_0^{(1)}(-1)^j$, and $h_0^{(1)}$ is obtained by numerical optimization. The resulting eigenvalues are shown in Fig. 2 (c). Most eigenvectors have increased eigenvalues except one, which provides significant improvement for inference, but the remaining one eigenvalue may still cause difficulty. This example shows that a single perturbation sometimes is not sufficient to obtain large eigenvalues for all eigenvectors. Effects of perturbation strongly depend on network structure, as illustrated in a complete analysis of three-node networks (*SI Appendix*, Supplementary text, Fig. S1−3). However, as FI is additive for independent samples, we can combine the information from many samples with different choices of local fields. In the geometric viewpoint, eigenvectors with small eigenvalues in the FI represent flat, singular valleys with diminishing second-order derivative near the maximum and, therefore, low local curvature in the likelihood landscape. Combining samples taken from different conditions is equivalent to adding these landscapes together. As the singular dimensions will differ with different perturbations, combining the landscapes can make the overall landscape well-behaved. In the Ising chain example, another field $h^{(2)}$ with $h_j^{(2)} = h_0^{(2)}\cos(\pi j/2)$ can be used to improve the eigenvalue on the previous degenerate direction. Neither $h^{(1)}$ nor $h^{(2)}$ alone improves eigenvalues in all eigenvectors, but the combination of the two perturbations improves all eigenvalues, as shown in Fig. 2 (c).

Iterative Bayesian Inference

To perform inference across a combination of different perturbations in a general setting, information obtained from different perturbations must be integrated. The inside argmax optimization problem in Eq. 4 is difficult, as the partition function $\mathcal{Z}$ in the objective function involves exponentially many terms. For optimization, the gradient of the log-likelihood has a closed-form expression

$$\frac{\partial \log \mathcal{L}}{\partial J_{ij}} = \langle s_i s_j \rangle^{\mathrm{o}} - \langle s_i s_j \rangle \ , \qquad [11]$$

where the $\langle s_i s_j \rangle^{\mathrm{o}}$ is the average over the observed samples, and $\langle s_i s_j \rangle$ is the average over the distribution generated by the current parameters. Even though exact evaluation still involves exponentially many terms, the gradient can be approximated by samples taken from Markov chain Monte Carlo sampling. In general, the inference cannot converge to the ground truth because of the accumulation of sampling error and the singularity of FI, and that is where perturbations can help.

Updating MLE with the latest round of samples can be viewed as maximizing a Bayesian posterior, using posterior of all previous samples as prior. The inference process not only finds the most probable parameters, but also updates the posterior probability of each parameter given the samples observed. The difficulty is that Bayesian inference can be computationally intractable if we need to compute and save the whole posterior each time. However, the gradient of log posterior can be computed by the recursive formula Eq. 14. In the equation, $\mathcal{L}_n$ is just the same likelihood function as in single round learning, while P_{n-1} is the posterior of the previous

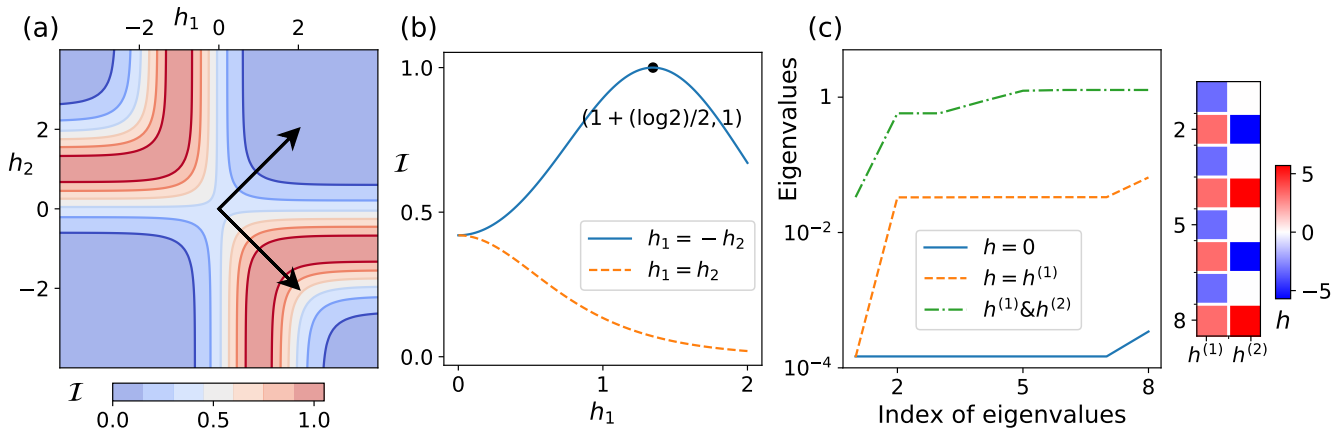

Fig. 2. Examples of two-node inference and Ising chain inference. The interaction $J = 1$ in (a)(b). (a) Landscape of the Fisher information $\mathcal{I}$ with different applied field perturbations h_1, h_2. Two arrows show the two principal-axis directions (eigenvectors of the Hessian matrix) of the saddle point at the origin, along which the FI is plotted in (b). (b) Fisher information as a function of h_1 when setting $h_1 = -h_2 \geq 0$ or $h_1 = h_2 \geq 0$. In the first case the maximum $\mathcal{I} = 1$ is achieved at $h_1 = J + \log\sqrt{2}$. (c) Eigenvalues of 1-dimensional ferromagnetic Ising model with $p = 8$, $J = 3$. When no field is applied, all eigenvalues are very small. When the field $h^{(1)}$ applied, all eigenvalues increase significantly except one. Combined with another field $h^{(2)}$, all eigenvalues are within a suitable range for precise inference.

$$P_n \equiv P(\boldsymbol{J}|\bigcup_{i=1}^{n}\{\boldsymbol{s}\}_i, \boldsymbol{h}_i) = \frac{P(\{\boldsymbol{s}\}_n|\boldsymbol{J}, \boldsymbol{h}_n)P(\boldsymbol{J}|\bigcup_{i=1}^{n-1}\{\boldsymbol{s}\}_i, \boldsymbol{h}_i)}{\sum_{\boldsymbol{J}} P(\{\boldsymbol{s}\}_n|\boldsymbol{J}, \boldsymbol{h}_n)P(\boldsymbol{J}|\bigcup_{i=1}^{n-1}\{\boldsymbol{s}\}_i, \boldsymbol{h}_i)} \quad [12]$$

$$\log P_n = \log \mathcal{L}_n + \log P_{n-1} - \log Z_n \quad [13]$$

$$\frac{\partial \log P_n}{\partial J_{ij}} = \langle s_i s_j \rangle_n^{\circ} - \langle s_i s_j \rangle_n + \frac{\partial \log P_{n-1}}{\partial J_{ij}} . \quad [14]$$

round. The normalizing factor Z_n does not contribute to the gradient. The computational cost of the Bayesian gradient only depends linearly on the number of learning rounds. The log likelihood is additive, and the final landscape defined by the posterior is the superposition of the landscapes for each individual perturbation.

Intuitively, our knowledge of the system will increase in this process, which can be proved by properties of FI. According to the Cramér-Rao bound, the ℓ_2 error of unbiased inference is bounded by $\mathrm{Tr}\,\mathcal{I}^{-1}$. By properties of positive semidefinite matrices, $A \succeq B \Leftrightarrow A^{-1} \preceq B^{-1}$. Therefore

$$\mathrm{Tr}(\mathcal{I}_1 + \mathcal{I}_2)^{-1} - \mathrm{Tr}\ \mathcal{I}_1^{-1} = \mathrm{Tr}((\mathcal{I}_1 + \mathcal{I}_2)^{-1} - \mathcal{I}_1^{-1}) \leq 0 . \quad [15]$$

Thus, the lower bound of the ℓ_2 error of the estimator decreases with more training rounds.

The choice of perturbation $\boldsymbol{h}$ is critical to improving the inference accuracy. As mentioned, $\mathrm{Tr}\,\mathcal{I}^{-1}$ serves as a asymptotic lower bound for the ℓ_2 error of any unbiased estimator. With $\boldsymbol{h}_i, i = 1, \ldots, n-1$ already set, the optimal choice of $\boldsymbol{h}_n$ in the n-th round is the solution of the optimization problem

$$\begin{aligned} \min_{\boldsymbol{h}_n} \quad & \mathrm{Tr}\ \mathcal{I}_n^{-1} \\ \text{s.t.} \quad & \mathcal{I}_i = \mathcal{I}_{i-1} + \mathcal{I}(\boldsymbol{J}, \boldsymbol{h}_i) \quad i = 1, \ldots, n \\ & \mathcal{I}_0 = 0 , \end{aligned} \quad [16]$$

where $\mathcal{I}_n$ is defined recursively. However, when using the method to uncover the structure of networks in applications, $\mathcal{I}(\boldsymbol{J}, \boldsymbol{h}_i)$ cannot be evaluated directly as $\boldsymbol{J}$ is unknown. So we need to approximate $\mathcal{I}(\boldsymbol{J}, \boldsymbol{h}_n)$ with our current estimate $\tilde{\boldsymbol{J}}$. For $i = 1, \ldots, n-1$, we already acquired samples from the real system, thus $\mathcal{I}_i$ can be approximated using the empirical average to replace the true distribution average in Eq. 6. The procedure runs in an iterative way between computation and experiments: new perturbations are designed based on previous samples and our resulting estimate $\tilde{\boldsymbol{J}}$. Each time with new samples taken from the system, solving the optimization Eq. 16 gives the most informative perturbation to execute in the next experiment. This framework could also be expanded to perform multiple new perturbations each time, with multiple $\boldsymbol{h}_n$ set as free decision variables each round. As the previous examples show, this optimization problem is highly complex and non-convex. In applications, we use a quasi-Newton method to find a reasonable choice of $\boldsymbol{h}$.

Results

Inference with oracle fields. First, we demonstrate that good perturbations can dramatically improve the precision of inference. Specifically, a medium-sized network can be deciphered with a few fields provided by an oracle that has a model of the underlying network. The oracle finds good perturbations by numerically optimizing Eq. 16 using the ground truth network, $\boldsymbol{J}$, and provides these perturbations to the inference procedure.

In many real systems, networks are composed of several communities or modules (23). One common form is a network that has activation inside each module and repression between different modules. For such systems, samples obtained from the natural condition are always inadequate to infer the exact position of these interactions. We performed our method on a 16-node network with three modules, which is amenable to

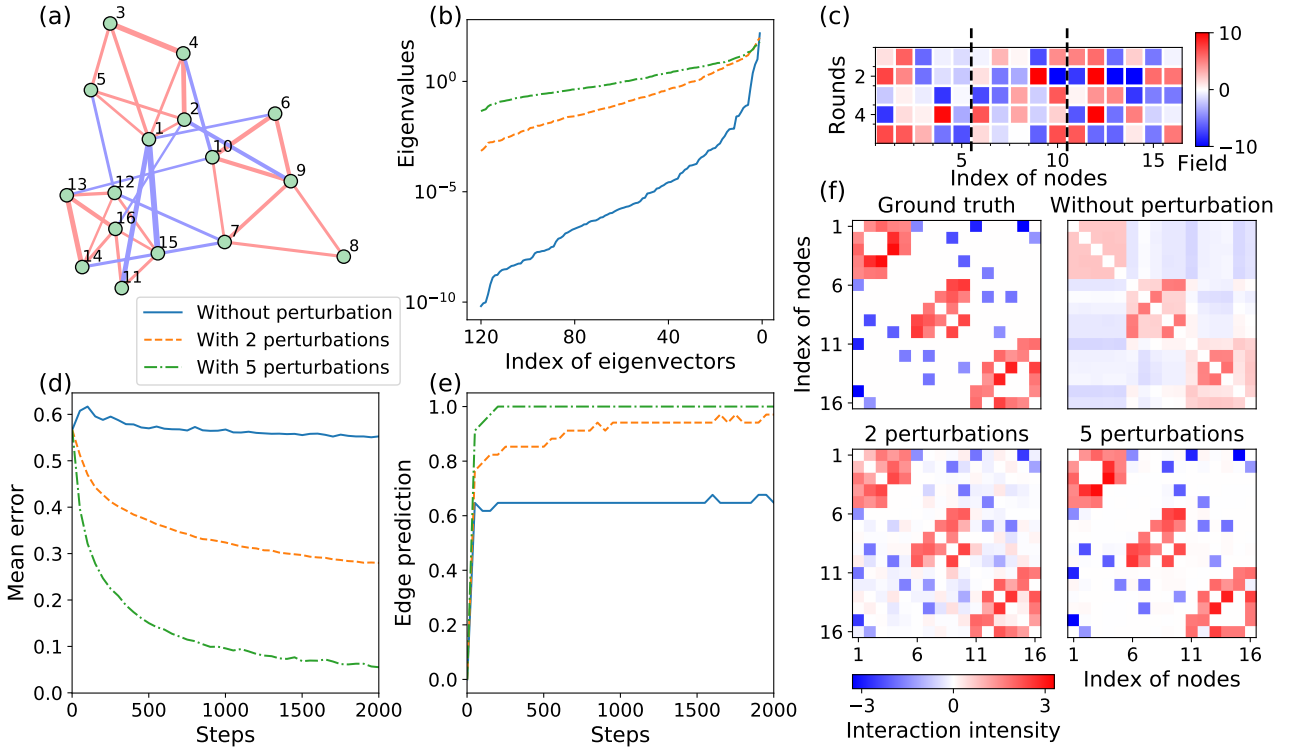

Fig. 3. Inference of a modular network with fields provided by an oracle. (a) The structure of the network to be inferred. Red edges represent $J_{ij} > 0$ and blue edges represents $J_{ij} < 0$. The width of each edge is proportional to $|J_{ij}|$. (b) Eigenvalue spectrum of Fisher information matrix of the inference problem without perturbation and with a given number of perturbations. Legend for (b)(d)(e) shown below panel (a). (c) Heatmap visualization of the applied perturbations across learning rounds. Perturbation magnitude indicated by color. In the first round no perturbation is applied so there are 5 perturbations in total. The black dashed lines separate different modules in the network, and the modular structure can also be seen in (a). (d) Mean estimation error of J_{ij} as a function of training steps. The total number of samples for each experiment is the same. (e) Edge prediction precision as a function of training steps. The strongest K edges in the prediction are compared with all the K edges of the real $\boldsymbol{J}$ to quantify the ratio of correct edges in the prediction. (f) Interaction matrix $\boldsymbol{J}$ and the estimate $\tilde{\boldsymbol{J}}$ with different number of perturbations are presented as heatmaps with $p \times p$ pixels.

numerical analysis while being large enough to model cases of interest. The structure of the network is shown in Fig. 3 (a), and the weights are set as random numbers to avoid special symmetry. As shown in Fig. 3 (b), some FI eigenvalues of the original inference problem are as small as 10^{-10}. We take 5×10^6 samples from the distribution each time, so there is no possibility to achieve accurate inference on the eigenvectors with 10^{-10} eigenvalues.

To demonstrate the existence of informative perturbations, $\boldsymbol{h}$, we perform numerical optimization of $\boldsymbol{h}$ with the true $\mathcal{I}$ and $\boldsymbol{J}$ in Eq. 16, and provide the resulting optimal field to the learning procedure as an oracle. The oracle fields are illustrated by a heatmap in Fig. 3 (c). After applying the field, the eigenvalues of FI are significantly increased by orders of magnitude. With only two perturbations, the smallest eigenvalue is $\sim 10^{-4}$, which is reasonable to infer with our sample size. We define two measures to quantify the improvement of inference after applying the perturbations. The mean estimation error is defined as $\sum_{i \neq j} |J_{ij} - \tilde{J}_{ij}|/n(n-1)$, the training curve of which is shown in Fig. 3 (d). Denote the number of edges in the true network as K. We define the edge prediction precision to be the normalized overlap between the most significant predicted K edges and the ground truth, as shown in Fig. 3 (e).

Without perturbation, the average prediction error of $\tilde{\boldsymbol{J}}$ does not decrease, as the improvement on correct edges is accompanied by false links in wrong edges. The edge-prediction results can also be visualized by the heatmap of links shown in Fig. 3 (f). Prediction without perturbation produces roughly all positive connections inside each module, and negative connections between modules. This phenomenon agrees with our intuition, that for strongly coupled networks we can only know the composition of modules, but not the exact interactions inside and between modules. With two perturbations, the mean prediction error decreased to around 0.3, which is 10% of the mean interaction strength. Also, the edge prediction precision increased from 0.6 to ~ 1 compared with the case of no perturbation. So we could learn the structure of the network qualitatively with two perturbations. Moreover, with five perturbations, we could obtain quantitative knowledge of the network. The mean edge prediction error goes down to 2% of the mean interaction strength, and prediction precision converges to 1 very quickly in the training, as shown in Fig. 3 (c)(e). The perturbations provided by the oracle do not have an obvious pattern. In general, the perturbations try to break the strong coupling inside the module by forcing

the nodes to have different values. Also, stronger fields are applied to nodes that have more links in general.

Inference with inferred fields. The oracle method cannot be applied in real systems, as the structure of the network is unknown and an oracle that provides good perturbations is generally unavailable. In real systems, to perform inference, we need to infer informative $\boldsymbol{h}$ using the empirical FI and our estimate $\tilde{\boldsymbol{J}}$ in Eq. 4. To validate that our active learning method is still effective to find good perturbations with empirical FI and $\tilde{\boldsymbol{J}}$, we perform the procedure in 49 randomly generated networks. Some examples of network structure are shown in Fig. 4 (a). The smallest eigenvalues of the original inference problem are around $10^{-7} - 10^{-10}$, therefore the inference is almost impossible with only 5×10^6 samples each round.

The results show that the perturbations discovered using estimation from data can still reveal network structure. After each round of sampling, the training results of the original problem and the perturbed problem are shown in Fig. 4 (b)(c). The mean training curves are shown in the opaque lines with standard deviation as error bar, and individual training curves are shown as transparent lines in the background. Without perturbation, after each round of sampling, the mean estimation error does not have observable change, and the prediction precision only improves slightly. In contrast, the training curves with the inferred perturbations improve significantly as we get more samples taken with different perturbations. For most networks, the edge prediction precision converges to 1 but with a different number of perturbations.

Even though inference in all networks is improved, the effect of 9 rounds of perturbation has some variation across networks. The final mean estimation error varies between $1\% - 10\%$ of the mean interaction strength. This is because our design of perturbation relies on the estimation quality of empirical FI and $\tilde{\boldsymbol{J}}$. For harder problems, our inference is less accurate, so the designed field is not as effective. The relation between the mean estimation error after 9 rounds of perturbation and the smallest eigenvalue of the inference problem without perturbation is shown in Fig. 4 (d). For inference without perturbation, the mean estimation error is insensitive to the smallest eigenvalues, as information of these directions is never captured under the given sample size. In comparison, the final error with perturbations decreases significantly with larger smallest eigenvalues. For smallest eigenvalues around 10^{-9}, even though the number of samples is not sufficient to find the network structure, certain directions are pinned down where the inference is hard, so that we can use additional perturbations to improve accuracy. Previously (Eq. 15), we showed that our knowledge of the system only increases with more perturbations, so we would expect convergence after enough rounds of perturbation.

In the process of finding a good $\boldsymbol{h}$, several approximations are made with some implicit assumptions. We argue that these approximations are valid in the sense of finding good perturbations. First, empirical FI is used instead of the true FI. By the theory of random matrices, the empirical FI converges to FI with increasing number of samples, and the convergence rate for different eigenvectors is proportional to the exponential of the eigenvalues (24). So we will have accurate FI estimation along those "easy" directions. Also, $\mathcal{I}_n$ is computed using the estimate $\tilde{\boldsymbol{J}}$ in place of $\boldsymbol{J}$. As FI is the expected Hessian of the log-likelihood function, by the theory of optimization, the convergence rates of the estimation is proportional to eigenvalues along different eigenvectors (25). In both cases, we will have accurate estimation along the eigenvectors whose eigenvalues are large, which is also supported by the numerical results (*SI Appendix*, Supplementary text, Fig. S4). These eigenvectors represent network components for which inference is accurate based on current samples, such as the positive interactions inside the module and negative interactions between modules. When designing new perturbations, we would like the new perturbation to reveal the information we have not obtained yet. Even though we do not have accurate knowledge for certain parts of a network, the inferred perturbations provide information that helps identify the directions along which our estimation is inadequate.

Discussion

In this paper, we developed a framework to rationally design and analyze perturbation experiments for parameter estimation. Our results show that perturbations designed to minimize the trace of the inverse of FI of the inference problem can provide significant improvements in both qualitative structure prediction and quantitative interaction strength estimation. Our framework combines statistical inference with active exploration, and thus mimics the scientific discovery process. Our method differs from traditional active learning methods, which typically select new samples with uncertainty-related criteria. Instead, we select perturbations that manipulate a given system to reveal information about the most uncertain properties. Also, compared with previous work on causal inference using Bayesian networks, our framework does not depend on specific properties of the model, but interprets the role of perturbation as "curving" previously singular dimensions. Many interesting statistical questions arise in this framework and need further exploration.

In practical situations where information about optimal perturbations is not available (Fig. 4), we use approximation in our optimization Eq. 16 to find a good choice of $\boldsymbol{h}$. Even though we have numerical evidence and empirical arguments to show that the discovered perturbations will still be informative under these approximations, strict analyses are still lacking. FI is widely recognized as a measure of the uncertainty in parametric inference, but the uncertainty in FI estimation is not as well-studied, which is essential in our case to know how good our proposed $\boldsymbol{h}$ will be. Moreover, the sampling complexity of such a learning scheme has not been established. Results on specific examples studied here show exponential improvement, but generalization to arbitrary networks needs more sophisticated analyses.

We demonstrate that good choice of $\boldsymbol{h}$ (using an oracle) yields dramatic improvements in inference in Fig. 3. In practice, as we only have the samples taken from the system itself, the best possible perturbation should be defined in the sense of posterior distribution on all possible $\boldsymbol{J}$. Even though oracle $\boldsymbol{h}$ enable more efficient inference depending on $\boldsymbol{J}$, trying to find such $\boldsymbol{h}$ is in some sense impossible as we do not have the required information in $\boldsymbol{J}$. We can use the expected decrease of $\mathrm{Tr}\,\mathcal{I}^{-1}$ on the posterior, or maximize the minimum improvements for a subset of the most likely $\boldsymbol{J}$.

Except for a few special cases, we find optimal fields numerically by finding fields that optimize the trace of the inverse of FI given the current estimate of network parameters. When

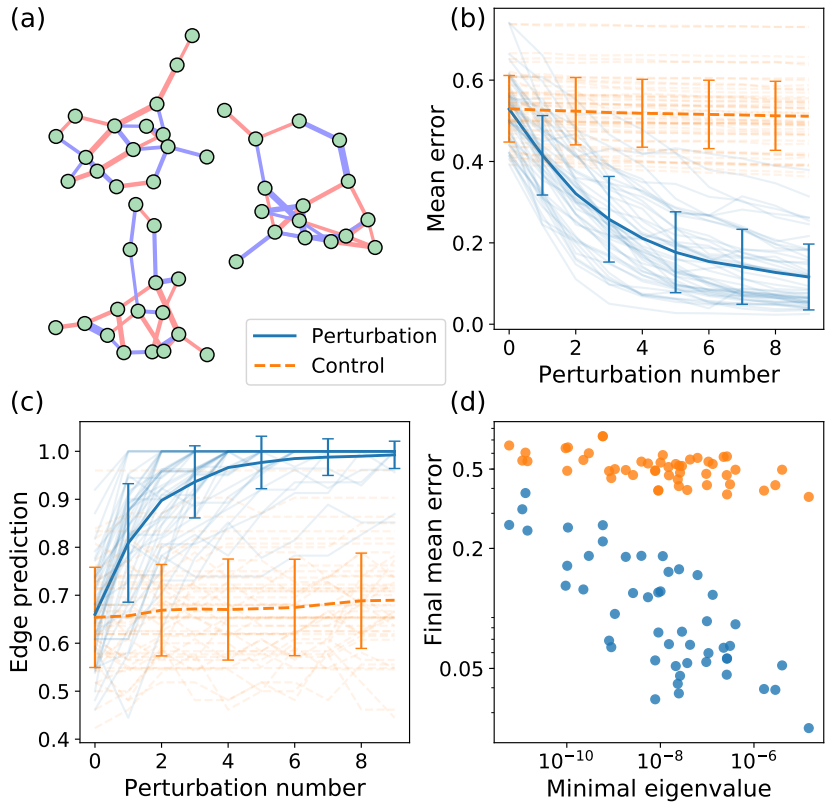

Fig. 4. Inference of random networks with inferred fields. (a) Examples of some 16-node random networks used in numerical experiments. (b) Mean estimation error after training is shown as a function of the number of applied perturbations. The opaque line and error bar represents the mean and standard deviation over all 49 tested random networks. The control group is set as taking the same extra number of samples but without perturbation. The transparent lines in the background show all trajectories for these networks. Legend for (b)(c)(d) shown below panel (a). (c) Edge prediction precision is shown as a function of the number of applied perturbations. The definition is the same as in Fig 3 (e). Definition of opaque and transparent lines is the same as (b). (d) Log-log plot between the final mean estimation error and the smallest eigenvalue of FI of the system without perturbation.

applied to large networks, the optimization might be computationally intractable, and it would be more efficient if we could design $\boldsymbol{h}$ directly from $\boldsymbol{J}$ and $\mathcal{I}$ without estimating the FI after hypothetical perturbation. Intuitively we would like the applied field to improve the probability of states that have not appeared before, as demonstrated in examples of three-node networks (*SI Appendix*, Supplementary text, Fig. S2−3). Preliminary results on finding $\boldsymbol{h}$ based on principal components analysis of the correlation matrix show some utility, but further investigation is required to make this approach practical.

All the above analyses are based on the framework we proposed where we have full control of the field term, and can apply any possible field to the studied network. Generally in applications, our ability to perturb the system might be more constrained. For example, perturbations might be constrained in magnitude or in ℓ_0 norm, which brings more challenges to theoretical analysis and optimization. Additionally, some central nodes could be essential to the proper function of the system and cannot be perturbed. Another possibility is that our control of $\boldsymbol{h}$ is imprecise, that the actual applied $\boldsymbol{h}$ includes a random component. Even though the illustrated example, spin networks, may not fully characterize the studied systems, the framework and procedure could be extended to other models of complex networks, such as Bayesian networks or dynamical systems. We believe that our framework provides an interesting approach to design and analyze perturbations in order to improve inference. The theoretical questions that arise in the process might provide new insights into statistical learning theory.

Materials and Methods

Numerical experiments were performed in MATLAB (26).

Learning of network parameters. For each round of learning, 5×10^6 examples are taken from the network by MCMC sampling. The optimization is performed by gradient ascent, and 5×10^3 samples are used in each step to estimate the gradient. The step size is chosen as $\eta = \lambda t^{-\alpha}$, where $\lambda = 0.1$, $\alpha \in [0.2, 0.5]$ depending on learning stages. To avoid over-fitting, ℓ_2 regularization is used during the training.

Generation of random networks. The random networks are generated by cutting off Gaussian random variables. Each edge is assigned a weight from the standard normal distribution, and we only keep the weights larger than 1.4 in magnitude. Then all remaining weights are rescaled to make their mean absolute value equal to 2.5.

Optimization of applied fields. For 16-node networks, accurate Fisher information was used in the computation. When computing the trace of the inverse, an identity matrix with 10^{-6} weight was added to avoid numerical instability. The optimization was performed by the optimization toolbox in MATLAB (26).

ACKNOWLEDGMENTS. The authors would like to thank Venkat Chandrasekaran and Andrew Stuart for influential discussions, and Yifan Chen for helpful suggestions. The authors would like to acknowledge support from the the Heritage Medical Research Institute (MT), the NIH (DP5 OD012194) (MT), and the Natural Sciences and Engineering Research Council (NSERC) Discovery Grant (DAS), and a Tier-II Canada Research Chair (DAS).

1. Le Novere N (2015) Quantitative and logic modelling of molecular and gene networks. *Nature Reviews Genetics* 16(3):146.
2. Hsu PD, Lander ES, Zhang F (2014) Development and applications of crispr-cas9 for genome engineering. *Cell* 157(6):1262–1278.
3. Dixit A, et al. (2016) Perturb-seq: dissecting molecular circuits with scalable single-cell rna profiling of pooled genetic screens. *Cell* 167(7):1853–1866.
4. Deisseroth K (2011) Optogenetics. *Nature methods* 8(1):26.

5. Paninski L (2005) Asymptotic theory of information-theoretic experimental design. *Neural Computation* 17(7):1480–1507.
6. Hyttinen A, Eberhardt F, Hoyer PO (2013) Experiment selection for causal discovery. *The Journal of Machine Learning Research* 14(1):3041–3071.
7. Molinelli EJ, et al. (2013) Perturbation biology: inferring signaling networks in cellular systems. *PLoS computational biology* 9(12):e1003290.
8. Ideker TE, THORSSONt V, Karp RM (1999) Discovery of regulatory interactions through perturbation: inference and experimental design in *Biocomputing 2000*. (World Scientific), pp. 305–316.
9. Murphy KP (2001) Active learning of causal bayes net structure, Technical report.
10. Tong S, Koller D (2001) Active learning for structure in bayesian networks in *International joint conference on artificial intelligence*. (Citeseer), Vol. 17, pp. 863–869.
11. He YB, Geng Z (2008) Active learning of causal networks with intervention experiments and optimal designs. *Journal of Machine Learning Research* 9(Nov):2523–2547.
12. Cho H, Berger B, Peng J (2016) Reconstructing causal biological networks through active learning. *PloS one* 11(3):e0150611.
13. Nguyen HC, Zecchina R, Berg J (2017) Inverse statistical problems: from the inverse ising problem to data science. *Advances in Physics* 66(3):197–261.
14. Marks DS, et al. (2011) Protein 3d structure computed from evolutionary sequence variation. *PloS one* 6(12):e28766.
15. Lezon TR, Banavar JR, Cieplak M, Maritan A, Fedoroff NV (2006) Using the principle of entropy maximization to infer genetic interaction networks from gene expression patterns. *Proceedings of the National Academy of Sciences* 103(50):19033–19038.
16. Cocco S, Leibler S, Monasson R (2009) Neuronal couplings between retinal ganglion cells inferred by efficient inverse statistical physics methods. *Proceedings of the National Academy of Sciences* 106(33):14058–14062.
17. Hinton GE, Osindero S, Teh YW (2006) A fast learning algorithm for deep belief nets. *Neural computation* 18(7):1527–1554.
18. Aurell E, Ekeberg M (2012) Inverse ising inference using all the data. *Physical review letters* 108(9):090201.
19. Vuffray M, Misra S, Lokhov A, Chertkov M (2016) Interaction screening: Efficient and sample-optimal learning of ising models in *Advances in Neural Information Processing Systems*. pp. 2595–2603.
20. Santhanam NP, Wainwright MJ (2012) Information-theoretic limits of selecting binary graphical models in high dimensions. *IEEE Trans. Information Theory* 58(7):4117–4134.
21. Montanari A, Pereira JA (2009) Which graphical models are difficult to learn? in *Advances in Neural Information Processing Systems*. pp. 1303–1311.
22. Watanabe S (2009) *Algebraic geometry and statistical learning theory*. (Cambridge University Press) Vol. 25.
23. Newman ME (2006) Modularity and community structure in networks. *Proceedings of the national academy of sciences* 103(23):8577–8582.
24. Tropp JA, , et al. (2015) An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning* 8(1-2):1–230.
25. Boyd S, Vandenberghe L (2004) *Convex optimization*. (Cambridge university press).
26. (R2018b) MATLAB, version 9.5.0.942161. The MathWorks Inc., Natick, MA, USA.

Supporting Information Text

Analytical solutions of specific models

Optimal Fisher information for two-node inference. In two-node case, the maximum of FI is achieved when $\langle s_1 s_2 \rangle = 0$, which requires

$$\frac{1}{\mathcal{Z}} \left(e^{J+h_1+h_2} + e^{J-h_1-h_2} - e^{-J-h_1+h_2} - e^{-J+h_1-h_2} \right) = 0 \,. \quad [1]$$

Solving the equation gives

$$h_2 = \frac{1}{2} \log \frac{1 - \exp(2J + 2h_1)}{\exp 2J - \exp 2h_1} \,. \quad [2]$$

Note that the equation is well-defined when $h_1 < J$ or $h_1 > J$, corresponding to the hyperbolic structure shown in main text.

Fisher Information of Ferromagnetic Ising Chain. The correlation $\langle s_i s_{i+1} \rangle$ and quadruple correlation $\langle s_i s_{i+1} s_j s_{j+1} \rangle$ can be computed by the methods of transfer matrix. The partition function without an external field can be written as

$$\mathcal{Z} = \sum_{s_1,\ldots,s_p} \exp(J \sum_{i=1}^{p} s_i s_{i+1}) = \operatorname{Tr} P^p \,, \quad [3]$$

where

$$P = \begin{bmatrix} e^J & e^{-J} \\ e^{-J} & e^J \end{bmatrix} \,. \quad [4]$$

The eigenvalues of transfer matrix are

$$\lambda_1 = e^J + e^{-J} \quad \lambda_2 = e^J - e^{-J} \,. \quad [5]$$

By symmetry of the Ising chain with periodic boundary conditions,

$$\begin{aligned} \langle s_i s_{i+1} \rangle = \langle s_1 s_2 \rangle &= \frac{1}{\mathcal{Z}} \sum_{s_1,\ldots,s_p} s_1 s_2 \exp(J \sum_{i=1}^{p} s_i s_{i+1}) \quad [6] \\ &= \frac{1}{\mathcal{Z}} \operatorname{Tr}(\frac{\partial P}{\partial J} P^{p-1}) = \frac{\operatorname{Tr} Q P^{p-1}}{\operatorname{Tr} P^p} = \frac{\lambda_1 \lambda_2^{p-1} + \lambda_2 \lambda_1^{p-1}}{\lambda_1^p + \lambda_2^p} \,, \end{aligned}$$

where Q is defined as

$$Q = \frac{\partial P}{\partial J} = \begin{bmatrix} e^J & -e^{-J} \\ -e^{-J} & e^J \end{bmatrix} \quad [7]$$

For the quadruple correlation function, noticing that

$$PQ = QP = \lambda_1 \lambda_2 \operatorname{diag}(1, 1) \,, \quad [8]$$

we have

$$\langle s_i s_{i+1} s_j s_{j+1} \rangle = \frac{1}{\mathcal{Z}} \operatorname{Tr} \left[P^{i-1} Q P^{j-i-1} Q P^{p-j} \right] = \frac{\lambda_1^2 \lambda_2^{p-2} + \lambda_2^2 \lambda_1^{p-2}}{\lambda_1^p + \lambda_2^p}, \quad i \neq j \,. \quad [9]$$

Then the series expansion at $e^J \to \infty$ of $I_{\{i,i+1\},\{j,j+1\}} = \langle s_i s_{i+1} s_j s_{j+1} \rangle - \langle s_i s_{i+1} \rangle \langle s_j s_{j+1} \rangle$ will give the result

$$\mathcal{I}_{\{i,i+1\}\{j,j+1\}} = \begin{cases} 4(p-1)\exp(-4J) & i = j \\ 4\exp(-4J) & i \neq j \end{cases} \,. \quad [10]$$

Additional examples of three-node inference

To help understand the mathematical framework and its implications, we provide a global analysis of three-node networks and detailed analysis of optimal perturbations for two specific three-node networks. In general, there are 7 different topologies of connected three-node networks up to a permutation, as shown in Fig. S1 (a), where the topology is defined as the signed edge connectivity. For three-node networks, the optimal perturbation can be found numerically by grid search of all possible directions, and so the networks provide a tractable set of examples in which we can explore the impact of perturbation on inference comprehensively.

Setting the absolute value of all interactions equal to 2, $\operatorname{Tr} \mathcal{I}^{-1}$ for each topology without perturbation and with one numerically optimal perturbation is shown for each topology in Fig. S1 (b). From this analysis, we can draw two general conclusions. First, the difficulty of inference depends, as represented by $\operatorname{Tr} \mathcal{I}^{-1}$, depends on network topology. Network 2, 4 and 6 are not fully connected and have smaller $\operatorname{Tr} \mathcal{I}^{-1}$ compared to other networks without perturbation. Therefore, these networks are intrinsically "easier" to learn by observation. Second, network topology also impacts on the optimal $\operatorname{Tr} \mathcal{I}^{-1}$ with perturbation. All of the not-fully-connected networks and network 3 and 7 achieve the lower bound 3 after 1 perturbation. Conversely, the perturbation only decrease $\operatorname{Tr} \mathcal{I}^{-1}$ of network 1 and 5 from 10^3 to 10^2. Therefore network 3 and 7 are "easy" to

infer with optimal perturbation, while network 1 and 5 are "hard" even with one optimal perturbation, which demonstrates the necessity of performing multiple rounds of perturbations for certain classes of networks. The different behavior of $\mathrm{Tr}\,\mathcal{I}^{-1}$ is determined by the energy landscape defined by $\boldsymbol{J}$, and we take network 3 and 1 for detailed analysis.

By the symmetry of the system, FI is the same if the sign of the field is flipped, so we can set $h_1 > 0$ without loss of generality. Then the applied field can be parametrized as

$$\boldsymbol{h} = |\boldsymbol{h}| \left[\sqrt{1 - h_2^2 - h_3^2}, h_2, h_3\right]^T , \tag{11}$$

where $|\boldsymbol{h}|$ is the Euclidean norm of $\boldsymbol{h}$. Given the direction of perturbation, the minimal of $\mathrm{Tr}\,\mathcal{I}^{-1}$ over $|\boldsymbol{h}|$ can be shown as a heatmap of $[h_2, h_3]$, as shown in Fig. S2 (b) and Fig. S3 (b).

For the network in Fig. S2 (a), the optimal FI achieves the lower bound $\mathrm{Tr}\,\mathcal{I}^{-1} = 3$, and the optimal perturbation is approximately $[2J, -2J, -2J]$. However, for the network in Fig. S3 (a), the minimum of $\mathrm{Tr}\,\mathcal{I}^{-1}$ is far larger than 3. As shown in Fig. S2 (b) and Fig. S3 (b), the direction of perturbation is crucial to the resulting optimal FI. The eigenvectors and eigenvalues without and with perturbation are shown in Fig. S2 (c) and Fig. S3 (c). These eigenvectors and eigenvalues can be interpreted in the network structure. For example, the smallest eigenvalue in inferring network 3 is the same as the signed edge connectivity. This is because increasing the edge intensity proportionally does not change the distribution much and is hard to determine. The effect of the perturbation on the distribution can be visualized by comparing the energy of all configurations, as shown in Fig. S2 (e) and Fig. S3 (e). Blue (orange) dots represent energies without (with) perturbation. The effect of perturbation is to create multiple low-energy states, in other words, to make some "informative" configurations have high probabilities to be observed.

Convergence of empirical Fisher information spectrum

In the main text, we argued that the empirical FI is a good estimation of FI on the eigenvectors with large eigenvalues. These arguments are supported by the numerical evidence shown in Fig. S4. As shown in Fig. S4 (a)(c)(e), the estimated FI gives a good estimation of eigenvalues of real FI when the eigenvalue is larger than around 10^{-5}. This is related to our sample size 5×10^6. The samples do not contain information about very small eigenvalues, so the corresponding eigenvalues in estimated FI are close to 0 up to the magnitude of numerical error. The estimation quality of eigenvectors can be quantified by computing the inner product between any eigenvectors in real FI and estimated FI. Examples of the absolute value of the inner product are shown in the inset plots of Fig. S4 (b)(d)(f). The eigenvectors are ranked by the magnitude of its eigenvalues, with an increasing order from left to right, and from top to bottom. There are two features of the estimation of eigenvectors. First, the estimation of eigenvectors is relatively precise for eigenvector with large eigenvalues. Second, the "mixing" of different eigenvectors happens between those with similar eigenvalues. Furthermore, the relation between estimation spread as a function of eigenvalues are shown in Fig. S4 (b)(d)(f). The spread is defined as the variance of a random variable X defined by

$$P_{v_i}(X = j) = \langle v_i, u_j \rangle^2 , \tag{12}$$

where v_i (w_j) is an eigenvector in the estimated (real) FI. The distribution is well-defined as v_i and w_j are taken from orthonormal basis. Generally, the spread is smaller with large eigenvalues. After 5 perturbations, most values of variance are around $10^0 \sim 10^1$, that is, mixing only happens to a few "neighbor" eigenvectors. The spread does not converge to 0 as there are near degenerate eigenvalues, and estimated FI only have the information of the spanned eigenspace by the corresponding eigenvectors. Even if the eigenspace estimation is accurate, the estimated eigenvectors are spanning over the whole eigenspace and thus different from the eigenvectors of the real FI. But the estimation of eigenvectors is still good, as visualized in the inset plot in Fig. S4 (f).

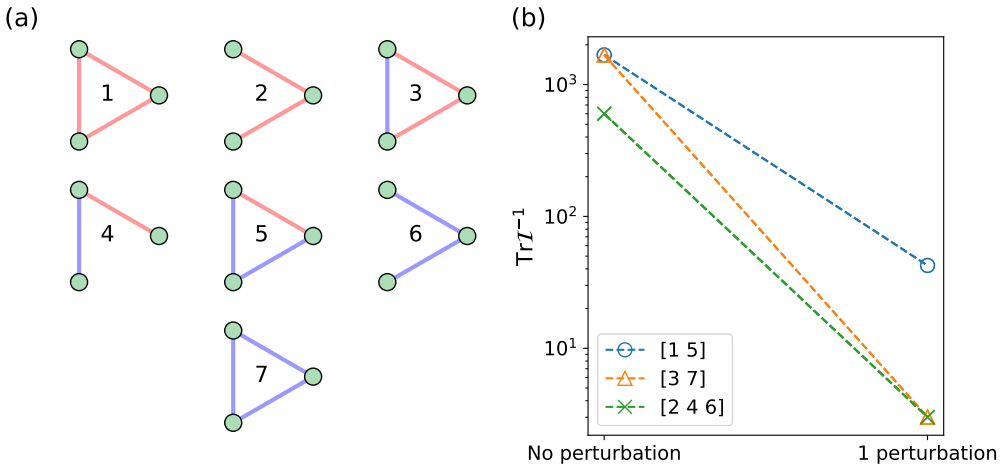

Fig. S1. The effect of perturbation on three-node networks. (a) All possible topologies of connected three-node network. Red (blue) edges represent positive (negative) connections. (b) Every magnitude of interaction strength is set to 2. Based on $\mathrm{Tr}\,\mathcal{I}^{-1}$ without and with 1 perturbation, these networks can be classified into three groups. The indexes of networks in each group are shown in the legend.

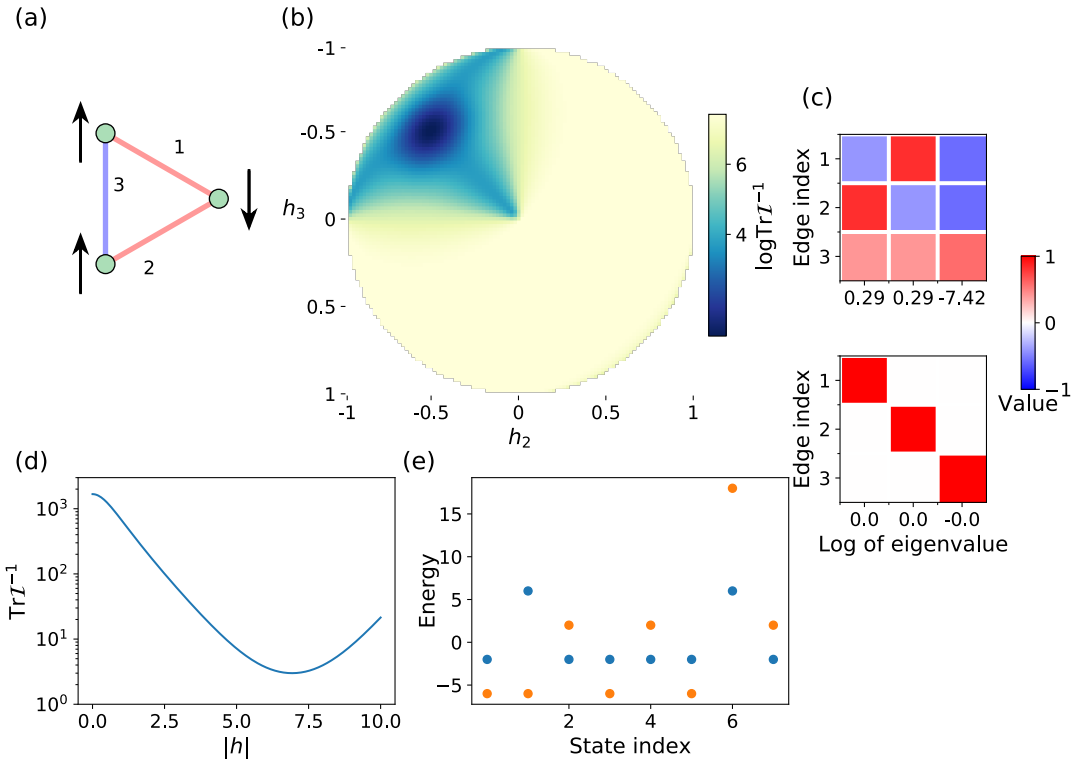

Fig. S2. Optimal perturbation of an "easy" three-node network (a) The structure of the network and the found optimal perturbation. Interaction strengths are set to 2 and the sign is indicated by the color as previously mentioned. The direction and length of the arrow represent the applied field on each node. (b) Minimal of $\mathrm{Tr}\,\mathcal{I}^{-1}$ as a function of the directional vector specified by h_2 and h_3 as in Eq.11. (c) Eigenvectors and eigenvalues of FI without and with perturbation. Each eigenvector is represented as a column in the heatmap, and the logarithm of the corresponding eigenvalue is shown below. (d) $\mathrm{Tr}\,\mathcal{I}^{-1}$ as a function of $|h|$ along the best perturbation direction. (e) The energy of each state without and with optimal perturbation. Blue (orange) dots are energies without (with) perturbation.

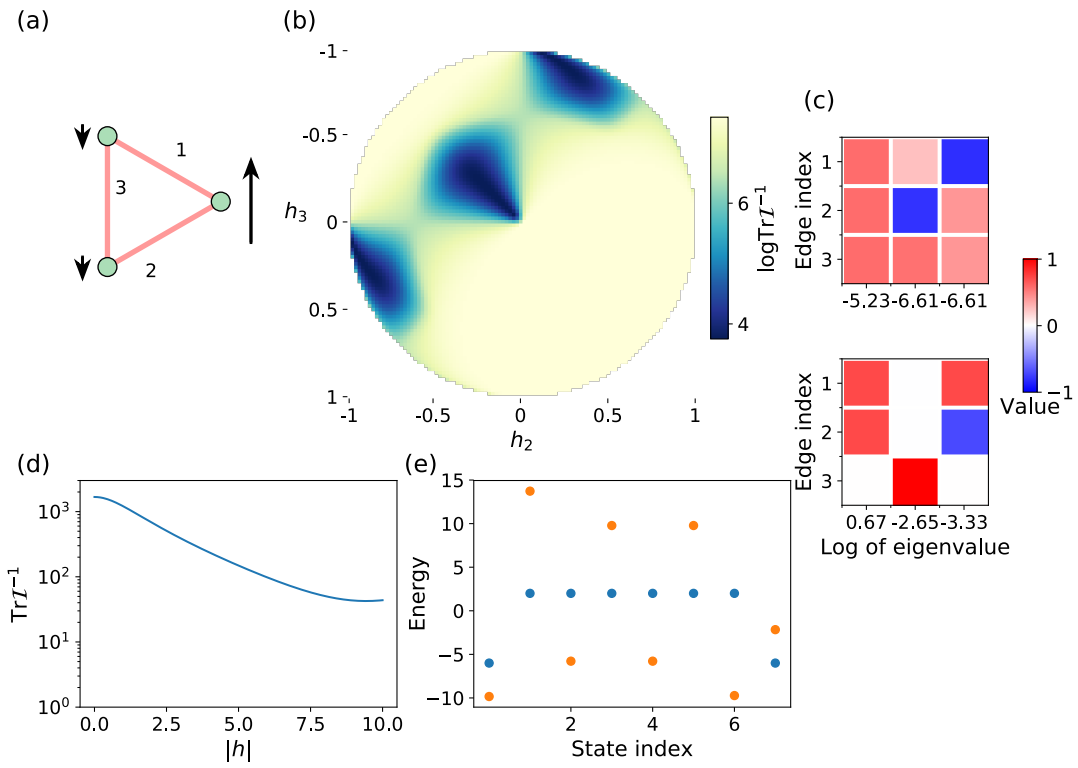

Fig. S3. Optimal perturbation of a "hard" three-node network. All other captions are the same as Fig. S2. ℓ_2-regularization is used in finding the optimal perturbation to avoid some singularity issues.

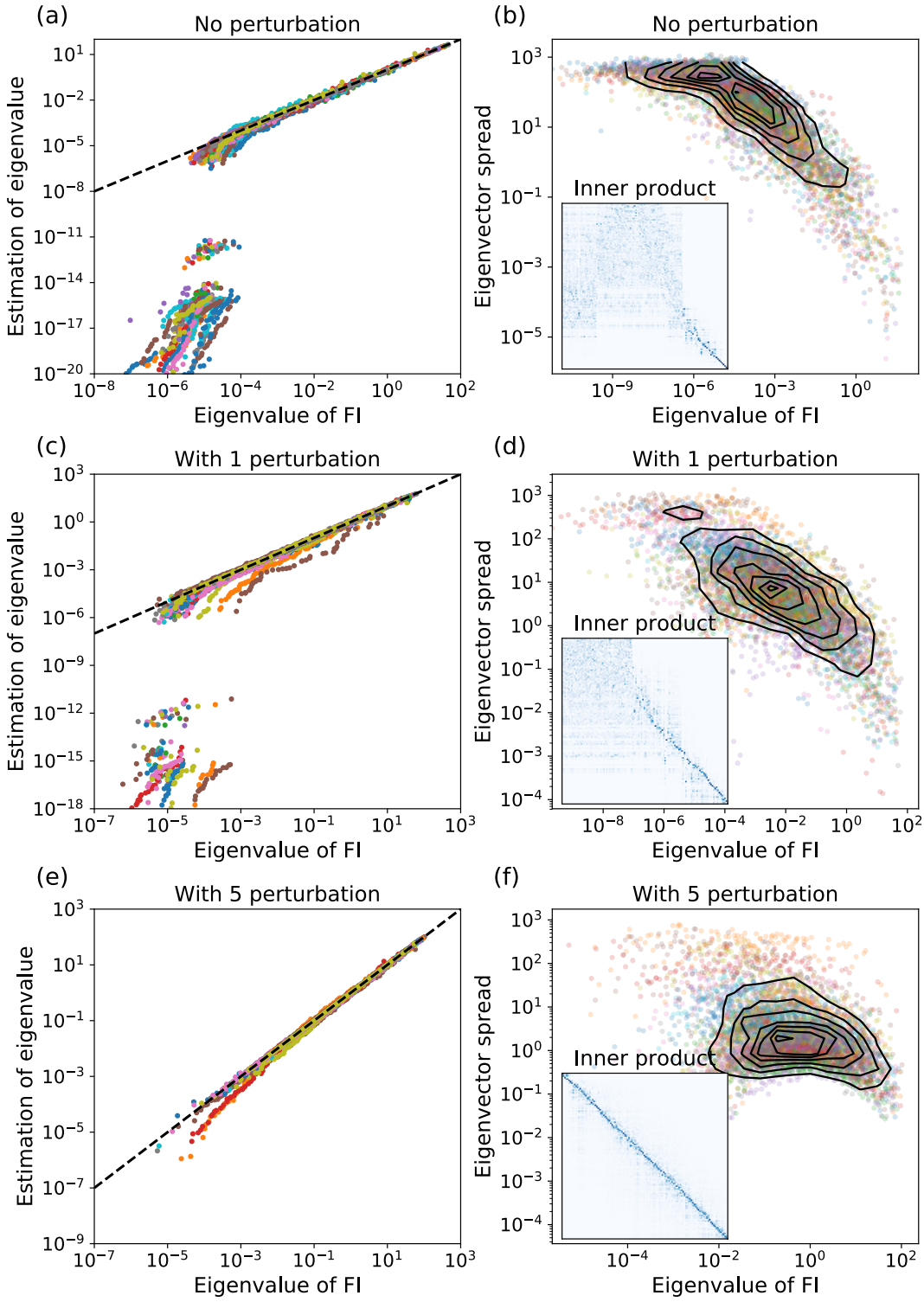

Fig. S4. Spectrum of estimated FI (a) The relation between ranked eigenvalues of real FI and estimated FI is shown as a scatter plot. Each color is a different network in the 49 tested random networks. The black dashed line is a reference $x = y$ line. (b) The spread of estimated eigenvectors is shown as a function of corresponding FI eigenvalues. Black lines are contours of the density of all points. The inset plot is an example of the absolute value of the inner product between eigenvectors of estimated FI and real FI. The corresponding eigenvalues of eigenvectors follow an increasing order from left to right, and from top to bottom. (c)(d) The same plot as (a)(b) for the estimated FI and real FI with one round of perturbation. (c)(d) The same plot as (a)(b) for the estimated FI and real FI with 5 rounds of perturbation.